COVID-19 Prediction Model

Machine Learning Project Documentation

Project Description
This project presents a complete machine learning pipeline for predicting COVID-19 infection and hospitalization risk using clinical and demographic data. The system includes data preprocessing, exploratory data analysis, model training, evaluation, optimization, and deployment preparation, following best practices in artificial intelligence and healthcare applications.

Faculty / University

- Arab Academy for Science, Technology and Maritime Transport

---

Course Code and Name

- CAI3101 – Introduction to Artificial Intelligence

Course Instructors

- **Dr. Mohamed Ali Abdel-Rabuh Hamouda**
- **Eng. Mohamed Moheb Abdel-Sattar Emara**

Submitted By

- **Seif Ebeid – ID: 231014746**
- **Abdullah Nagy – ID: 231004881**

---

1. Project Overview

This project presents a complete machine learning system for predicting COVID-19 infection and hospitalization risk using clinical and demographic data. The system implements an end-to-end pipeline that includes data cleaning, feature engineering, exploratory analysis, model training, evaluation, optimization, and deployment preparation.

Multiple machine learning algorithms are trained and compared to identify the most reliable and generalizable model for medical decision support. The project is designed for educational and research purposes, with a strong focus on medical validity, bias reduction, and model robustness.

---

2. Dataset Information

The dataset used in this project is a COVID-19 clinical dataset containing anonymized patient medical records.

- **Source**: COVID-19 clinical patient dataset **Link**
- **Original Size**: Approximately 400,000 rows and 24 columns (sampled for performance)
- **Final Dataset Size**: Variable, depending on medical filtering and validation steps
- **Target Variable**:
  - `covid` (Binary Classification)
  - 1 = COVID Positive
  - 0 = COVID Negative

Features Used

The final model relies only on medically relevant and pre-diagnosis features.

Demographic Feature

- HOSPITALIZED

Clinical Status

Age

Medical Condition Indicators

- USMER
- INTUBED
- PNEUMONIA
- PREGNANT
- DIABETES
- COPD
- ASTHMA
- HIPERTENSION
- OTHER_DISEASE
- CARDIOVASCULAR
- RENAL_CHRONIC
- OBESITY
- TOBACCO
- INMSUPR
- ICU

3. Data Cleaning and Preprocessing Pipeline

A strict data cleaning strategy was applied to ensure medical correctness, prevent data leakage, and reduce bias.

3.1 Data Leakage Prevention

The `DATE_DIED` column was removed because it represents post-outcome information that would not be available at prediction time. Keeping this feature would artificially inflate model performance.

3.2 Bias Reduction

The `SEX` column was removed to avoid demographic bias and ensure that predictions rely on clinical and medical indicators rather than personal attributes.

3.3 Target Variable Construction

A new binary target variable, `covid`, was created from the `CLASIFFICATION_FINAL` column:

- Values 1, 2, and 3 were mapped to COVID Positive (1)
- Values 4, 5, 6, and 7 were mapped to COVID Negative (0)

Records with invalid classification values were removed. After target creation, the original classification column was dropped.

3.4 Hospitalization Encoding

The `PATIENT_TYPE` column was renamed to `HOSPITALIZED` and re-encoded as follows:

- 0 = Outpatient
- 1 = Hospitalized

3.5 Binary Feature Standardization

All medical condition features were standardized into binary format:

- 1 → Yes / Condition Present
- 2 → No / Condition Absent

This ensured consistency across all categorical medical features.

3.6 Age Validation

Age values were converted to numeric format using coercion. Only valid ages between 0 and 120 years were retained.

3.7 Final Quality Control

- The original classification column was removed
- All binary columns were verified to contain only 0 or 1
- Records with invalid or inconsistent values were excluded

---

4. Exploratory Data Analysis (EDA)

Exploratory analysis was performed to understand data distributions, feature relationships, and potential predictive patterns.

4.1 Target Distribution

- Count plots were used to visualize COVID-positive vs COVID-negative cases
- Pie charts illustrated class proportions
- Statistical summaries confirmed class imbalance levels

4.2 Correlation Analysis

A correlation heatmap was generated to analyze relationships between features and the target variable. Feature ranking helped identify the strongest predictors of COVID infection.

4.3 Feature Distribution Analysis

Binary medical features were visualized individually and in grouped plots to assess prevalence and data completeness.

4.4 Age Analysis

Age distributions were compared between COVID-positive and COVID-negative patients using histograms, box plots, and descriptive statistics.

---

5. Machine Learning Models

Several models were trained to compare performance across different algorithm families.

Models Implemented

1. Logistic Regression
2. Decision Tree (Regularized)
3. Random Forest (Regularized)
4. K-Nearest Neighbors
5. Gradient Boosting
6. Gradient Boosting (Hyperparameter Tuned)
7. Voting Ensemble (GB + RF + Logistic Regression)

---

6. Feature Engineering

Feature Scaling

StandardScaler was applied to normalize numerical features for models sensitive to feature magnitude.

- Applied to: Logistic Regression, KNN, Voting Ensemble
- Benefit: Improved convergence and performance
- Performance Gain: Approximately 3–5% accuracy improvement

---

7. Model Evaluation Strategy

Train–Test Split

- 80% training, 20% testing
- Stratified split to preserve class distribution
- Fixed random state for reproducibility

Cross-Validation

- 5-fold cross-validation for general comparison
- 10-fold stratified cross-validation for the best model

Evaluation Metrics

- Accuracy
- Precision
- Recall (Sensitivity)
- F1-Score
- Specificity
- ROC-AUC

---

8. Advanced Validation Techniques

Learning Curves

Used to detect overfitting and underfitting by comparing training and validation performance across different dataset sizes.

ROC Curve Analysis

ROC curves and AUC scores were used to evaluate the model's ability to distinguish between positive and negative cases.

Confusion Matrix

Confusion matrices were analyzed with special focus on false negatives, which are critical in medical diagnosis.

Feature Importance

Feature importance analysis was performed using the best-performing model to identify key clinical predictors.

---

9. Optimization Techniques

Regularization

Model complexity was controlled using depth limits, minimum sample constraints, class balancing, and reduced ensemble sizes.

Hyperparameter Tuning

RandomizedSearchCV was used for efficient hyperparameter optimization, achieving performance improvements while maintaining short training times.

Ensemble Learning

A soft-voting ensemble combined multiple models to reduce individual model errors and improve stability.

---

10. Model Performance Summary

Typically observed performance:

- Accuracy: 60–75%
- Precision: 60–75%
- Recall: 65–80%
- F1-Score: 65–75%
- ROC-AUC: 0.65–0.80
- Overfitting Gap: Less than 5%

---

11. Conclusion

This project demonstrates a complete and medically responsible machine learning pipeline for COVID-19 prediction. Through careful data cleaning, bias reduction, model comparison, and validation, the system achieves reliable performance suitable for educational and research applications while respecting healthcare constraints.